DOCUMENT RESUME

ED 063 578 CS 000 006

AUTHOR Harris, Albert J.; Jacobson, Milton D.
TITLE Natural Language Computerized Comparison of Word List Content.
PUB DATE Apr 72
NOTE 11p.; Paper presented at the Annual Convention of American Educational Research Assn. (Chicago, April 3-7, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Automatic Indexing; *Computational Linguistics; Computer Programs; *Information Processing; *Language Research; *Structural Analysis; Structural Linguistics; Word Frequency; *Word Lists

ABSTRACT

The development of a computerized system of word analysis in order to compare and compile word lists is outlined. It is suggested that a computerized system would be an efficient way of comparing word lists for such elements as content (according to criteria of range, scope, and form of words), obsolescence, levels of difficulty, number of words, length of words, frequency of words appearing in reading materials, and construction of words (for example singular-plural and verb forms). Comparison of word lists can in turn lead to compilation of new lists based on specific requirements for various purposes. The program can also be utilized to research word associations, to score responses to programmed material, and to determine the comprehensibility of textual passages. (The process of computerizing information about words is described, possible uses for the program are suggested. A sample comparison of four word lists is given. Tables of data and references are included. (AL)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

Natural Language Computerized Comparison of Word List Content

Albert J. Harris
City University of New York

Milton D. Jacobson
University of Virginia

PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY
Albert J. Harris
& Milton D. Jacobson
TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."

ED 063578

The development of computerized analysis of verbal data provides a means by which the content of word lists can be quickly and comprehensively compared. Because of the importance of word lists to the development and standardization of the vocabulary of instructional materials, it is useful to compare and contrast word lists. This study was conducted in order to develop a computer program which would print out a comparison of word lists that would reveal overlap between lists, words unique to one list or another, and numerical level of reading proficiency assigned to the words in each of the lists.

A computer program capable of comparison of word list content seems useful for a variety of reasons. Most obvious is facilitation of comparison of word list content according to criteria of range, scope, or form of words which should be included. A quick, mechanized listing operation allows one to evaluate the differences in the vocabularies represented by two or more lists. A more subtle application might be the comparison of lists and the materials constructed with them in order to identify differences created by the passage of time.

CS 000 006

Some of the lists in widespread use today were developed as many as fifty years ago. A computerized comparison procedure allows one to evaluate the differences between old lists and

modern ones according to criteria of obsolescence in vocabulary. In effect, the process of aging can be isolated and identified, making the evaluation of the usefulness of old lists and the materials which they were used to construct a feasible task. As new lists are developed, their content can be compared, allowing users to evaluate the relative usefulness of one or another.

The procedure used to enable an automated comparison of word list content involved the punching of several lists onto IBM cards, then programming the computer to sort the words, compare them for correspondence, check for correspondence or variation in level assignment, and print out the results in verbal form. The lists compared are the Harris-Jacobson Basic Elementary Reading Vocabularies (1), the Dale List of 3,000 Words, (2), the Botel List (3), and the Taylor List for grades 1-8 and grades 9-13, (4). The words could have been punched either one to a card, with level information punched into a defined column of the card, or sequentially, separated by commas or spaces and followed by level information. With a slight modification, the program could deal with the data when it was punched in the latter form, thus reducing the bulk of the data.

The computer programming can be broken into two stages. The first stage receives and stores the raw data of the lists, automatically alphabetizing the words. This stage of the program forms a file constituting a single list of the words contained in all lists, in effect merging the lists to be compared. Every

word contained in the lists is recorded once in alphabetical order. Each word is accompanied by a mask 96 columns long, allowing the recording of 96 pieces of information for each word, such as the lists in which it appears. These columns could be alotted so as to record level assignment made by the compilers of the lists. For instance, if the Harris-Jacobson list is stratified into six levels, six columns in the mask could record the H-J list, each bit denoting the level in which the word appears. The next group of bits could be alotted to the next list, broken down according to its assigned levels, and so on. The file composed by this first stage of the program incorporates facilities for generating new information, for updating, or for correction of the existing data.

The second stage of the program reads through the file compiled by the first stage, and prints and tallies the merged lists. This printer stage of the program inputs a list of potential titles to be sought in the mask of the stage-one file, checks the columns for the requisite information, and prints the words with the appropriate titles. A characteristic of this stage of the program is that it utilizes both fixed-field and floating-field editing. The fixed-field editing can be stipulated to cover a section of the file mask corresponding to a criterion list to which the other lists are compared. In the print-out, this area will be filled only with information pertaining to the criterion list. If information is not supplied, if a word does not appear in the criterion list, that space on the print-out

will be blank. In contrast, the compared lists are edited in a floating-field. If a piece of information is not supplied, the next piece supplied will in effect slide over to occupy its space. The result is a listing with all the words contained in all the word lists appearing in alphabetical order along the left margin. Next is a space in which the appearance or absence of the word in the master list can be noted. To the right the comparison lists in which the word appears are shown. The print thus records the unique words of each list, the words which appear in more than one list, and where they are matched. Level information for each word is also printed if such information is provided by the compilers of the list. This print-out can then easily be read, and the nature of matched and unmatched words can be observed.

In addition to the print-out of the merged and compared lists, the program tallies information about the results, such as the number of words in both of two lists, the number of words in one list not in the other, the number of matched words which have been assigned to the same level by both compilers, or similarly, different levels. Further, the program can print out a list of matched words without unmatched words, or the unmatched words from either list without the matches.

The data for the study consisted of four word lists. The first was the Harris-Jacobson Basic Elementary Reading Vocabulary recently developed by Albert J. Harris and Milton D. Jacobson.(1)

The H-J computer list for this study includes both the Harris-Jacobson 7,612 root words and 9,237 inflected forms, totalling 16,849 entries. This list was compared to three other word lists: the Dale list of 3,000 common words developed by Edgar Dale,(2) the Botel Bucks County list of 1,185 common words compiled by Morton Botel,(3) and the EDL vocabulary compiled by Stanford Taylor and others (4). The EDL vocabulary was broken into two sublists which were compared independently, one for levels 1-8 and one for levels 9-13.

The results of the comparison are shown in Table I.

See Page10 for Table I.

The two bottom rows of Table I are probably the most informative. Of the 2,946 words in the Dale List, 2,744 or 93 percent also appear in the Harris-Jacobson List. Of the 3,266 words in the Botel List (including inflected forms), 3,095 or 94 percent are also in the Harris-Jacobson List. Thus the overlapping among these three lists is quite high. The degree of overlapping with the two Taylor lists is lower. Of the 6,714 Taylor words for grades one through eight, 5,473 or 81 percent are also in the Harris-Jacobson List. This is not a surprising result, since the Harris-Jacobson List stops at sixth grade and the Taylor List includes words for grades seven and eight. The Taylor high school list shows still less overlapping, since only 179 of 2,426 Taylor secondary words are in the Harris-Jacobson List.

While these raw tallies are interesting, the printed comparison of the lists provides a means of discovering the more specific differences in the lists. The effects of aging, for example, are evident in the comparison of the Dale list and the Harris-Jacobson list.

The Dale list is a list of nearly 3,000 familiar words widely used in estimating the readability of reading materials. The words were listed by Dale if 80 percent of fourth graders who were questioned said they knew them (2). The Harris-Jacobson list was developed from a computerized word count applied to 14 basal series of widely-used elementary instructional materials totalling 127 books. It is a more comprehensive list than the Dale list, and includes a Core List, an Additional List, and subject-matter vocabularies. It is stratified into six grade levels, but in the four lowest levels of the Core List there should be a basis of comparison with the Dale familiar words. The words unique to one list or the other reveal evolution in vocabulary which shows the effect of the passing of time on readability-oriented word list.

The words in the Dale list not in H-J include the following words which seem obsolete or of diministed frequency of use now: afar, apiece, bedbug, bookkeeper, bran, buttermilk, candlestick, christen, codfish, cooper, fib, fret, goody, henhouse, jig, lard, lass, lice, overalls, reap, schoolmaster, sleigh, snuff, trolley, washtub. Conversely, the Harris-Jacobson contains the following words which have come into common use since the Dale list was

developed: TV [level 1], elevator [2], tractor [2], traffic [2], battery [3], camera [3], detective [3], experiment [3], helicopter [3], strike [3], astronaut [4], bargain [4], committee [4], concrete [4], hamburger [4], satellite [4]. Vocabulary evolves as new scientific terms come into general use, as current events bring words to forefront positions in newspapers and conversation, and as public attitudes change, allowing previously obscure words to come into more common use. The effect of these changes on word lists can be readily observed in the printed side-by-side comparison of lists.

In addition to such content-analysis comparision of word lists, the computerized comparison procedure allows a quick evaluation of the explanatory factors for the differences between word lists. One can easily observe patterns among the unique words of either list which reveal construction criteria of the lists which distinguish them, such as word endings and forms of compound words. Some large differences in the sizes of lists can be discovered to be due to the fact that the compilers of one list chose to include all the variants of a word as separate entries, while the compilers of the other list chose to list only the root word or its most common variants. One list may contain hyphenated words and the other may not, one list may have included proper nouns, and so on.

Comparison between the Harris-Jacobson list and the Botel list of 1,185 words (approximately 3,000 with variants) revealed contrasts caused by differences in construction criteria rather

than vocabulary. Analyzing the words unique to the Botel list revealed only four root words: berry, excite, fairground, and linesman. These words did not attain the frequency required to appear on the Harris-Jacobson list.

The other differences are due largely to criteria for inclusion used by the lists' compilers. For instance, the Botel list included the words Indian and Christmas which were excluded from the Harris-Jacobson list because they are proper nouns. Botel also included plurals which occurred too infrequently to be included by Harris and Jacobson, such as bedrooms, buses, postmen, lads, lighthouses, neckties and schoolrooms. Here, the criteria for inclusion varied: the Botel list included plurals for most nouns, at the same level as the singular, the Harris-Jacobson list evaluated plurals according to the same criteria as singular nouns--if the frequency pattern was sufficient, the word was included. As a result, plurals are usually included at a higher level than corresponding singulars, or not at all. Similar differences in criteria caused Botel to include variants of verbs excluded from the Harris-Jacobson list, such as "eater," "prizing," "welcoming."

Because the Botel list assigns levels to its constituent words, comparison with the stratification of the Harris-Jacobson list was possible. Of the words which were matched between the lists, approximately 1,700 were assigned the same level and approximately 1,400 were assigned different levels by the two lists' compilers. The words could be examined to determine

whether the words given the same level can be typed in contrast to those given differing level assignments or to determine whether there is a pattern of up-levelling or down-levelling of words between the lists.

Differences in compilation criteria seem to be revealed in the comparison of the EDL vocabulary for grades 1-8 and the Harris-Jacobson list. Of the approximately 1,200 words unique to the EDL list, 14 are assigned levels 4 or under by the EDL compilers, 65 are level 5, 132 are level 6, the remainder are levels 7 and 8. These figures would indicate that the differences between the lists occur at borderline frequencies, where a word may just meet the EDL criteria and just miss the Harris-Jacobson criteria.

The implementation of the program developed in this study will be of significant value to researchers desiring comparative statistical data regarding word list vocabularies, as it will enable lists to be compared in a variety of ways quickly and usefully. It may also be useful in analyzing word associations, utilizing Cureton's adaptation of the Kuder-Richardson Formula 20 in the development of associative norms. Other uses are in scoring responses to programmed material and in determining the comprehensibility of textual passages.

TABLE I

COMPARISON OF THE HARRIS-JACOBSON BASIC ELEMENTARY READING VOCABULARY WITH FOUR OTHER WORD LISTS

	LIST BEING COMPARED			
	Dale List	Botel List	Taylor (1-8)	Taylor (9-13)
Total Number of Words in Harris-Jacobson List	16,849	16,849	16,849	16,849
Total Number of Words in Comparison List	2,946	3,266+	6,714	2,426
Number of Words in Harris-Jacobson That Are Not in Comparison List	14,105	13,754	11,376	16,670
Number of Words in Both Lists	2,744	3,095	5,473	179
Number of Words in Comparison Not in Harris-Jacobson	202	171	1,241	2,247

*Harris and Jacobson, Basic Elementary Reading Vocabularies
Of the 16,849 entries, 7,612 are root words in the published lists and 9,237 are inflected forms not printed as separate entries.
+Basically 1,185 words. When separate entries are made for each variant form it consists of 3,266 words (example: beat, beats, beating).

NOTES

1. Harris, Albert, and Milton Jacobson, Basic Elementary Reading Vocabularies. New York, Macmillan, 1972.

2. Dale, Edgar and Jeanne S. Chall, "A Formula for Predicting Readability: Instructions, "Educational Research Bulletin, Vol. XXVII, No. 2, February 17, 1948.

3. Botel, Morton, Botel Predicting Readability Levels, Chicago: Follett, 1962.

4. Taylor, Stanford E., Helen Frackenpohl, and Catherine E. White, A Revised Core Vocabulary: A Basic Vocabulary for Grades 1-8, and Advanced Vocabulary for Grades 9-13. Huntington, New York: McGraw-Hill, Educational Development Laboratories, 1969.